予測に生かす統計学

上越教育大学大学院学校教育研究科・学校教育学系
准教授　奥村太一

はじめに

統計学には，大きく記述，推測，予測の３つの役割がある。記述とは，得られたデータについて平均や標準偏差，相関係数といった統計的指標（統計量）を算出したり，度数分布表やヒストグラム，散布図などを作成したりすることで標本の特徴をまとめることを指す。推測とは，こうした記述の結果をもとにして母集団における様相を探ること，具体的には母平均や母標準偏差，母相関係数の値を推定によって見積もったり，それに関する仮説の真偽について検定によって意思決定したりすることを指す。一般的に，科学とは理論やモデルを構築し実験や観察を通じて得られたデータからその正しさについて検証するという営みであるから，そこで統計学が果たす役割は主に記述と推測である。しかしながら，理論にせよモデルにせよ適切に集められたデータからそれが支持されればその確実性は高まるものの，それが正しいことを証明し切ることはできない。実際，科学的に正しいと信じられてきた理論が新たな理論の出現によって置き換えられたり更新されたりすることは，ニュートン力学から相対性理論へ，さらには量子力学へという現代物理学の目まぐるしい進展を見ても明らかである。極端な言い方をするならば，科学研究の枠組みの中で記述と推測のサイクルを繰り返すことは，覚めることのない夢を見続けているようなものである。

統計学が科学研究にとどまらず広く一般社会で使われるようになったのは，予測という第３の役割によるところが大きい。予測とは，得られたデータをもとに将来起こりうる個々の現象について議論することを指す。気圧配置と降水量の関係について議論するのが推測だとすれば，今日の気圧配置から明日雨が降るか議論するのが予測だということになる。永遠に未知の値を追い求める推測とは異なり，予測は当たるか外れるかを実際に目の当たりにすることになるため，その結果が社会に及ぼす影響はより大きい。また，結果をモデルの改良に生かして予測の精度を高めることも自ずと求められることになる。実際，私たちのほとんどは気象の厳密な仕組みよりも明日傘は必要かということに関心があるのだし，そうした社会的要請を踏まえて日々工夫が重ねられることで天気予報の精度も上がり続けているのである[1]。

回帰と予測

イギリスの統計学者フランシス・ゴルトンによって「回帰」という概念が提唱されたことは以前に述べた[2]。これは，データの利用を相関関係の考察という静的な段階から，具体的な値の予測という動的な段階へと推し進める非常に重要な貢献であった。回帰の最も単純な例は，ゴルトンが取り組んだように父親の身長 X からその子供の成人時身長 Y を直線

$$\hat{Y} = \alpha + \beta X$$

によって予測するというものである。ただし，$\hat{Y}$ は Y の予測値であり，実測値と予測値のずれ $Y - \hat{Y}\left(= e\right)$ を残差という。変数 X と Y について大きさ N のデータが得られていれば，この残差の２乗和を最小にするような切片 $\hat{\alpha}$ と傾き $\hat{\beta}$ を求めることができる。また，残差 e が独立に平均 0，分散 σ^2 の正規分布に従うと仮定すれば（$e \sim N\left(0,\ \sigma^2\right)$），例えば傾き $\hat{\beta}$ については

$$\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}\right)$$

となることが知られているので[3]，

---

[1] https://www.data.jma.go.jp/fcd/yoho/kensho/yohohyoka_top.html
[2] http://www.jikkyo.co.jp/contents/download/9992658085
[3] X_i は i 番目の X の値，$\bar{X}$ は X の標本平均を表す。

$$P\left(\hat{\beta}-1.96\times\sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}\leqq\beta\leqq\hat{\beta}+1.96\times\sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}\right)=0.95$$

を利用して傾きβに対する信頼度95%の信頼区間を算出することができる[4]。これが推測の手続きであり，$\hat{\alpha}$や$\hat{\beta}$が求められていれば新たに値X_0が得られた場合のYの予測値を$\hat{Y}_0=\hat{\alpha}+\hat{\beta}X_0$と求めることができる。さらに，推測の結果を利用すれば，予測に関わる不確実性の程度を

$$P\left(\hat{Y}_0-1.96\sigma\sqrt{1+\frac{1}{N}+\frac{\left(X_0-\bar{X}\right)^2}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}\leqq Y\leqq\hat{Y}_0+1.96\sigma\sqrt{1+\frac{1}{N}+\frac{\left(X_0-\bar{X}\right)^2}{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}}\right)=0.95$$

のように予測区間と呼ばれる幅によって評価することもできる[5]。このような一連の分析を**回帰分析**と呼ぶ。

回帰分析にもとづいて予測を行うためには，切片や傾き等に関する情報をあらかじめデータから得ておく必要がある。このように，予測のためにあらかじめ用意しておくデータを「教師データ」といい，教師データにもとづいてこうした情報を整備することを「教師あり学習」などという。予測値$\hat{Y}_0$に対する実測値Y_0が得られれば，X_0とY_0を教師データに加えることで切片や傾きの値を更新することができる。つまり，予測と結果のフィードバックを繰り返すことで学習が進み，予測の正確さを高めたり不確実性を正確に見積もったりすることができるようになるということである。こうしたサイクルのことを，**統計的学習**とか機械学習などという。回帰分析は，統計的学習の最も基本的なモデルの1つである。

説明変数の追加

回帰分析は，例えば知能検査の結果（IQ）から標準学力検査の偏差値を予測するとか，種子の大きさから成長後の収穫量を予測するといったように，様々な現象の予測に応用することができる。しかし，予測区間に残差の標準偏差σが含まれていることからわかるように，予測の正確さ自体は統計的学習を進めれば直ちに高まるとは限らない。生徒の学力は知能だけでなく学習に対する態度や家庭環境などによっても左右されるであろうし，収穫量は日照時間や土壌環境にも依存するだろう。このように個人差や個体差を規定する要因が複数あるのであれば，それらを組み合わせて予測を行ったほうがより結果は正確になるはずである。

そこで，説明変数を$\left(X_1,\ X_2,\ \cdots,\ X_p\right)$のように複数用意し，

$$\hat{Y}=\alpha+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$$

のように予測することを考える。このように，複数の説明変数を含む場合を**重回帰分析**と呼ぶ。回帰分析と言った場合は通常この重回帰分析を指し，前節のように説明変数を1つしか持たないものは**単回帰分析**と呼んで区別することが多い。説明変数が複数になった場合も基本的な考え方は単回帰分析の場合と同様で，残差の2乗和を最小とするような切片と傾きをデータから求め，予測に用いることになる。

なお，説明変数は必ずしも大小が議論できるような量的なものである必要はない。例えば，成人後の身長を予測するのであれば，父親の身長だけでなく本人の性別も説明変数に入れた方が正確になるだろう。この場合，例えば女性$=0$, 男性$=1$のように性別を便宜的に数値に置き換えたものを用いることが一般的である。このように数値化されたものをダミー変数という。ダミー変数は数値であっても単に属性を分類するだけの機能しかなく，何らかの大小関係を反映したものではない。仮に，X_1を父親の身長，X_2を性別（ダミー変数）として身長Yを予測するとすると，女性は$X_2=0$，男性は$X_2=1$であることから

$$\text{女性：}\hat{Y}=\alpha+\beta_1X_1$$
$$\text{男性：}\hat{Y}=\alpha+\beta_1X_1+\beta_2$$

となり，傾きβ_2は，同じ身長の父親を持つ男女について平均身長の差を取ったものに相当することがわかる。

---

[4] ただし，実際は残差の母分散σ^2は未知であるから，これをデータから推定した値で置き換えたものがt分布と呼ばれる確率分布に従うことを利用して信頼区間を求めることになる。

[5] 信頼区間と同様，未知の母分散σ^2はデータから推定した値で置き換え，t分布を利用して求めることになる。

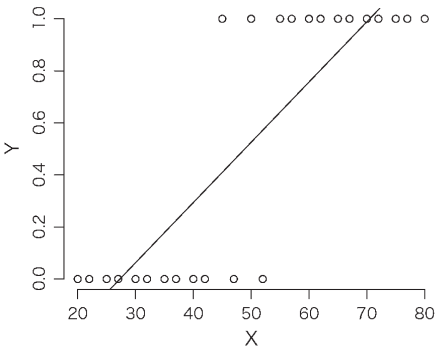
現在，コンビニなどで買物をしてポイントカードを提示した場合，私たちの購買記録はカードを発行する際に登録した性別や年齢などの個人情報と合わせて，瞬時にデータ化されている。このようなデータを POS（point of sales）**データ**と呼ぶ。いつどのような属性の人がどの商品を購入したのかについて日々膨大なデータが蓄積されているので，企業にとってはこれらをうまく活用することで新商品の開発や市場への投入時期，広告戦略を効果的に組み合わせて売上を伸ばすことが期待できる。

2値の現象を予測する

回帰分析では，用意された説明変数の線形結合によって Y が表されると仮定されている。統計学においては，このように現象を定式化したものを**モデル**と言い，モデルを作ることを**モデリング**などと言う。いかに現象に応じた適切なモデルを作れるかが予測の正確さを左右する大きな決め手であり，分析者の経験やセンスが問われるところである。

ここで，現実社会で予測の対象となるものは，必ずしも身長や収穫量，売上のように大小や多寡が議論できるものとは限らない。例えば，ある属性を持つ人が実際に特定の商品を購入するかどうかや，送られてきたメールを受信者が迷惑メールと判断するかどうかなど，何らかの事象が生起する／しないというように2値の現象を予測したい場合も少なからず存在する。通常の回帰分析では，例えば傾き β が正であれば，X の値が大きくなれば Y はそれに合わせていくらでも大きい値として予測されるので（負であれば逆），このままでは仮に Y を 1 と 0 にダミー変数化したとしても適切な予測ができないことは明らかである（右段上図）。

このような2値の現象を適切にモデリングするため，次のように回帰分析を拡張する。まず，$Y=1$ となる（事象が生起する）確率を p とおく。そのうえで，



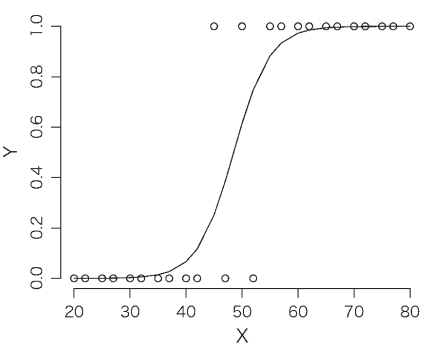
$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right)=\alpha+\beta X$$

のように予測することを考える。左辺は生起確率と非生起確率の比（オッズ）の対数を取ったものであり，これをロジットと呼ぶ。この式は

$$\hat{p}=\frac{1}{1+e^{-(\alpha+\beta X)}}$$

と変形できるから，事象の生起確率 p を S 字型のロジスティック曲線によって表していることになる。結果変数 Y については，独立に試行数 1，確率 p の二項分布（これをベルヌイ分布と呼ぶ）に従って発生する（$Y\sim B(1,p)$）と考えればよい。あとは，X と Y に関するデータが得られればそれに最もよく合う切片と傾きを求めることができる[6]。

下図は，先ほどと同じ2値データに対してロジスティック曲線を当てはめた様子を示したものである。直線を当てはめた場合に比べて，予測値と実測値のずれは明らかに小さくなっていることが見て取れる。回帰分析をこのように拡張したものを**ロジスティック回帰分析**と呼ぶ。例えば X が



[6] ただし，この場合は最小2乗法ではなく，得られたデータを発生する確率が最大となるような α や β の値を探索的に求める最尤法と呼ばれる方法が用いられるのが一般的である。

模擬試験の偏差値，Yが大学入試の合否であるならば，本番の試験で合格を勝ち取る確率が80%となるために模擬試験でどれくらいの成績をおさめておかなくてはいけないのか事前に把握することができる。また，曲線の立ち上がりが急であればそれだけ合否は正確に予測できることから，出題者にとっては模擬試験の品質評価や改善につながる情報を得られることだろう。私たちがネット通販で買い物をするとすぐにおすすめ商品が提示されるのも基本的にはこれと同じ仕組みである。過去の膨大な購買データから次にカートに入れられそうな商品を予測して提示し，それに対する私たちの反応がまたフィードバックされて予測の正確さを高めるシステムが運用されているのである。

線形モデルからの脱却

回帰分析であれロジスティック回帰分析であれ，説明変数Xの増加とともにYやpの予測値が単調に増加（もしくは減少）していくと考えていることに変わりはない。しかし，実際の現象はもっと複雑であって，年齢層が高ければ高いほど購入金額も高いとか，本文に“!”記号が多く含まれているほど迷惑メールである確率が高いといったような単純な関係ばかりではない。もちろん，Xの2乗や3乗の項を説明変数として投入するといった対処も考えられるが，結局はこうした高次の項と線形関係にあることを想定しているに過ぎない。予測の精度を高めるためには，むしろ「説明変数の重み付き和」という制約を解除してやったほうが場合によっては得策である。計算機の性能が飛躍的に向上したことに伴い，以前は扱うことができなかった切片や傾きが結果変数と非線形関係にあるとするモデル[7]も容易に扱うことができるようになってきた。それと期を同じくして，これまでは人間が行ってきた説明変数の取捨選択や関数形の決定に至るまで，様々な候補から最適なものを選び出し，モデルを自動的に構築するアルゴリズムまで整えられるようになっている。

[7] もはや「切片」や「傾き」という呼び方すら適切ではないかもしれないが。

おわりに

商品を提示したり迷惑メールを選別したりといった実用的な場面においては，とにかく予測が当たればそれで目的は達成されるのであって，どのように予測すれば正確なのか，なぜ予測が成功するのかといったことをいちいち人間が詮索し理解する必要はない。統計学が科学研究を越え，日常世界でも一般的なツールとして使われるようになった結果，私たちの生活は便利になった一方，意思決定の自動化が進み，その中身はブラックボックス化しているとも言える。2016年に，IBMの人工知能（AI）ワトソンが患者の病名を特定して有効な治療法を提案したという報道がなされた[8]。人工知能も基本的な仕組みは統計的学習であり，ワトソンも膨大な医学研究結果を教師データとして学習することでこのような正確な予測を下すことが可能となったのである。しかし，「なぜ」ワトソンがそのような診断に至ったのかは誰にもわからない。

今や，統計学は思考の道具から実用の道具へと大きく変貌を遂げつつあるように見える。将来的に，人間は考える機会を奪われ，AIの予測に隷属する存在に成り下がるのだろうか？私はそうは思わない。いくら正確に予測がなされたとしても，その意味を理解し評価することは人間にしかできないからである。「模擬試験でこの成績ではまず志望校に合格する見込みはない」という予測がはじき出されたときに，そもそも進学は諦めたほうが良いと捉えるか，志望先をより自分に合ったものに変えるチャンスと捉えるか，進学を一年延長して再度じっくり勉強に取り組もうと考えるかは人それぞれである。こうした行動の選択には個々人の信念や価値観が多分に反映されるのであって，唯一絶対の正解があるわけではない。確率・統計を学習する醍醐味も，同じ事実をもとに様々な解釈を考えうるというところにある。その意義は人間の尊厳と共にこれからも決して廃れるものではないと信じている。

[8] https://www.nikkei.com/article/DGXLZO05697850U6A800C1000000/